

A rate balance principle and its application to queueing models

Binyamin Oz*, Ivo Adan[†] and Moshe Haviv*

October 12, 2015

Abstract

We introduce a rate balance principle for general (not necessarily Markovian) stochastic processes. Special attention is given to processes with birth and death like transitions, for which it is shown that for any state i , the rate of two consecutive transitions from $i - 1$ to $i + 1$, coincides with the corresponding rate from $i + 1$ to $i - 1$. This observation appears to be useful in deriving well-known, as well as new, results for the Mn/Gn/1 and G/Mn/1 queueing systems, such as a recursion on the conditional distributions of the residual service times (in the former model) and of the residual inter-arrival times (in the latter one), given the queue length.

1 Introduction

Consider a (not necessarily Markovian) stochastic process with \mathcal{S} as its state-space and partition it in three sets: \mathcal{D} , \mathcal{M} , and \mathcal{U} . We define an *up path segment* as a path segment which commences in some state in \mathcal{D} , ends in \mathcal{U} , and uses as intermediate states (if any) only states in \mathcal{M} . In a similar fashion, we define a *down path segment*. We show that in any (finite) time interval, the number of up and down path segments differ by at most one. This implies that their steady state rates (if they exist) coincide. The case $\mathcal{M} = \emptyset$ is the well known balance principle between the sets \mathcal{D} and \mathcal{U} . Another special case, on which we dwell in the sequel, are processes with birth and death like transitions. In this case, with \mathcal{M} consisting of the single

*Department of Statistics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem

[†]Department of Industrial Engineering, Technische Universiteit Eindhoven

state i , what we get is that the rate of two consecutive transitions from $i - 1$ to $i + 1$ (namely, transitions that avoid getting again into state $i - 1$ before reaching $i + 1$), equals the rate of corresponding ones from $i + 1$ to $i - 1$. Note, however, that this result does not extend to states further than two transitions away from each other. Through a number of examples, we show the usefulness of this special case in deriving known, and also new, results in single server queues. In particular, it leads to an alternative derivation for the limiting probabilities in M/G/1 and G/M/c queues and to less known results on the residual service and inter-arrival times given the queue lengths.

Section 2 states the main result which we call the *rate balance principle* (RBP). A few examples are given as well. Section 3 presents some preliminaries on the distribution of the residual of a random variable given it is larger than an independent exponentially distributed random variable. Section 4 shows how the limiting probabilities of the G/M/1 queue can be derived using the RBP. This is repeated for the G/Mn/1 queue in Section 5. We also derive a recursion on the distribution function of the residual of the inter-arrival times at departure instances given the queue length. In Section 6 we derive the corresponding results for the Mn/Gn/1 queue, in which case the residuals of the service times at arrival instances are of concern. To the best of our knowledge, the recursions on the residuals of the conditional inter-arrival times and service times are new. Finally, Section 7 concludes.

2 Rate balance principle

Let $X = \{X(t), t \geq 0\}$ be a continuous-time stochastic process with state space \mathcal{S} . As in Section 3.2 of [6], we assume that \mathcal{S} is a Polish (complete separable metric) space, with the Borel σ -field $\mathcal{B}(\mathcal{S})$, and that X is right continuous with left-hand limits. Let \mathcal{D} and \mathcal{U} be two non-empty and disjoint subsets of the state space, $\emptyset \subsetneq \mathcal{D}, \mathcal{U} \in \mathcal{B}(\mathcal{S})$, and let $\mathcal{M} = \mathcal{S} \setminus (\mathcal{D} \cup \mathcal{U})$. We are interested in two types of path segments of this process. The first type are path segments that begin with a state in \mathcal{D} , end with a state in \mathcal{U} , and any other states in the segments (if any) are in \mathcal{M} . The second type are path segments that begin in \mathcal{U} , end in \mathcal{D} and any other states (if any) are in \mathcal{M} . We refer to such path segments as \mathcal{U} -segments and \mathcal{D} -segments, respectively. More formally, let $\{I_n^{\mathcal{U}}, n \geq 1\}$ and $\{I_n^{\mathcal{D}}, n \geq 1\}$ be two point processes indicating the time instances where X gets into \mathcal{U} and \mathcal{D} , respectively, i.e.,

$$I_n^{\mathcal{V}} = \inf\{t > I_{n-1}^{\mathcal{V}} | X(t^-) \notin \mathcal{V}, X(t) \in \mathcal{V}\}, \quad \mathcal{V} \in \{\mathcal{U}, \mathcal{D}\}, \quad n \geq 1,$$

where $I_0^\mathcal{V} := 0$, $\mathcal{V} \in \{\mathcal{U}, \mathcal{D}\}$. Assume that the partition of \mathcal{S} is such that a.s. X gets into \mathcal{U} and \mathcal{D} infinitely often in $[0, \infty)$, but at most finitely often in every finite time interval $[0, t)$, $t \geq 0$.

A \mathcal{U} -segment end point is the first time instant that X gets into \mathcal{U} , after getting into \mathcal{D} , and a \mathcal{D} -segment end point is defined symmetrically. Formally, the counting process of \mathcal{U} -segments and \mathcal{D} -segments are defined by,

$$N^\mathcal{U}(t) = \#\{k | I_k^\mathcal{U} \leq t, \exists m \text{ s.t. } I_{k-1}^\mathcal{U} < I_m^\mathcal{D} < I_k^\mathcal{U}\}$$

and

$$N^\mathcal{D}(t) = \#\{k | I_k^\mathcal{D} \leq t, \exists m \text{ s.t. } I_{k-1}^\mathcal{D} < I_m^\mathcal{U} < I_k^\mathcal{D}\},$$

respectively. We denote by $\{T_n^\mathcal{V}, n \geq 1\}$, the point process associated with $N^\mathcal{V}$, $\mathcal{V} \in \{\mathcal{U}, \mathcal{D}\}$.

The following theorem states that the steady state rates at which the two types of path segments occur are equal.

Theorem 1 *The following limits, if they exist, are equal:*

$$\lim_{t \rightarrow \infty} \frac{N^\mathcal{U}(t)}{t} = \lim_{t \rightarrow \infty} \frac{N^\mathcal{D}(t)}{t}.$$

Proof: Suppose that $I_k^\mathcal{U}$ is a \mathcal{U} -segment end point. We argue that the previous segment end point is a \mathcal{D} -segment end point. Let $m^* = \min\{m | I_{k-1}^\mathcal{U} < I_m^\mathcal{D} < I_k^\mathcal{U}\}$. Note that m^* exists since it is the minimum of a finite, non-empty set. Observe that $I_{m^*-1}^\mathcal{D} < I_{k-1}^\mathcal{U} < I_{m^*}^\mathcal{D}$ and hence, $I_{m^*}^\mathcal{D}$ is a \mathcal{D} -segment end point. The fact that the previous \mathcal{U} -segment end point is less than or equal to $I_{k-1}^\mathcal{U}$ completes the argument. Reversing the roles of \mathcal{U} and \mathcal{D} gives a symmetric argument.

The above implies that the two types of path segments occur alternately and hence, $N^\mathcal{U}(t)$ and $N^\mathcal{D}(t)$ differ at most by 1 for all $t \geq 0$ and thus

$$\lim_{t \rightarrow \infty} \frac{N^\mathcal{U}(t) - N^\mathcal{D}(t)}{t} = 0,$$

which completes the proof. ■

Practically speaking, in order to use RBP, one should consider sets \mathcal{D} and \mathcal{U} such that \mathcal{U} -segments and \mathcal{D} -segments have a simple structure. For example, consider a partition $\mathcal{D}, \mathcal{U}, \mathcal{M}$, and consider the following directed graph: the set of vertices is \mathcal{M} , and the set of edges is $\{(i, j) : i, j \in \mathcal{M} \text{ and } i \rightarrow j \text{ is a possible transition}\}$. If this graph is acyclic, then for each $d \in \mathcal{D}$ and each $u \in \mathcal{U}$ there are finitely many path segments emanating from d and

ending in u (and vice versa). We give a couple of examples for such choice of \mathcal{D} and \mathcal{U} below.

Example 2.1 Rate-Balance Equation

This example is based on Theorem 3.7 in [6]. Let $A \in \mathcal{B}(\mathcal{S})$ and let $\mathcal{D} = A$ and $\mathcal{U} = A^c$ (so $\mathcal{M} = \emptyset$). Of course, Theorem 1 implies that the rate of transitions from A to A^c , i.e., the rate of transitions out of A , equals the rate of transitions from A^c to A , i.e., the rate of transitions into the set A . In case $\mathcal{S} = \mathbb{Z}$ and $A = \{k : k \in \mathbb{Z}, k \leq \ell\}$ for some $\ell \in \mathbb{Z}$, we get the classical *level crossing* argument.

Example 2.2 Two-Step Transitions (TST)

Consider a process with state space $\mathcal{S} = \mathbb{Z}^+$ where transitions are of size 1, e.g., the number of customers in a queueing system where customers arrive one by one and are served one at a time. For $n \geq 1$, let $\mathcal{D} = \{k : 0 \leq k < n\}$ and $\mathcal{U} = \{k : k > n\}$. Here, \mathcal{U} -segments have the following form: they begin with a transition from state $n-1$ to state n , and end with a transition from state n to state $n+1$, i.e., two consecutive up transitions. We refer to such path segments as n -two-step up transitions. Symmetrically, \mathcal{D} -segments have the form of two consecutive down transitions, from state $n+1$ to state n , and then from state n to state $n-1$. We refer to such path segments as n -two-step down transitions. Theorem 1 implies that the rates of n -two-step up transitions and n -two-step down transitions are equal. In that case, as in the general case, the reason for this equality is that the two transitions occur alternately. Suppose that an n -two-step up transition just occurred. That means that the process is currently in state $n+1$. In order for another n -two-step up transition to occur, the process must visit state $n-1$ first, and for that to happen, an n -two-step down transition must occur. Figure 1 shows an example of a sample path where 3-two-step transitions are marked with a solid line. Note that this result does not extend to three (or more) consecutive up (or down) transitions. For example, the rate of immediate transition from $n-1$ to n and then immediately to $n+1$ and $n+2$, does not coincide with the corresponding rate from $n+2$ to $n-1$.

3 Preliminaries

Let Y_λ and X be two independent generic random variables such that $Y_\lambda \sim \exp(\lambda)$, X is nonnegative and the cumulative distribution function (CDF), Laplace-Stieltjes transform (LST) and mean of X are denoted by F , F^* ,

$$\lambda \frac{F^*(\lambda)}{1 - F^*(\lambda)} = h(0) =: \gamma. \quad (6)$$

Solving for $F^*(\lambda)$ gives

$$F^*(\lambda) = \frac{\gamma}{\lambda + \gamma}. \quad (7)$$

Finally, (7) and (4) uniquely define the LST F^* as

$$F^*(s) = H^*(s) \frac{\lambda - s}{\lambda} \frac{\lambda}{\lambda + \gamma} + \frac{\gamma}{\lambda + \gamma} \quad (8)$$

■

Remark 3.1 Let H be such that $h(0)$ equals zero or infinity. Equation (6) implies that $F^*(\lambda)$ equals zero or one, respectively, for any F such that $H = D_{\lambda, F}$. Observe that $F^*(\lambda) = \int_0^\infty e^{-\lambda x} dF(x)$ is the probability that a nonnegative random variable X with CDF F is less than an independent exponential random variable with rate λ . Hence, $F^*(\lambda)$ being equal to zero or one contradicts the assumption $P(X \in (0, \infty)) > 0$.

4 The G/M/1 queueing model

In this section we derive some well-known results on the G/M/1 queue using the rate balance principle. Consider a G/M/1 queue where G , G^* , and $1/\lambda$ are the CDF, LST, and mean of the inter-arrival times, respectively. Service times are exponential with rate μ . Assume that the system is stable, so $\lambda < \mu$, and denote the steady state probability of having n customers in the system just before an arrival, or just after a departure instance by a_n . Denote by π_n , $n \geq 0$, the steady state probability of having n customers in the system at an arbitrary instance, and let R_n and R_n^* denote the CDF and LST, respectively, of the steady state distribution of the residual inter-arrival time at a departure instance, conditioned on the queue length n .

Theorem 3 *For $n \geq 0$, R_n^* is not a function of n . In other words, the residual inter arrival time and the number of customers in the queue at departure instances are independent.*

Proof: Since service times are memoryless, the queue length distribution is insensitive to the service regime, as long as it is work conserving and non-anticipating. Hence we assume w.l.g. that the service regime is Last Come First Served with preemption (LCFS-PR) and consider such a system at

departure instances. The number of customers just after departure equals the number of customers that the departing customer saw upon her arrival. Observe that this number is independent of any random variable solely defined by the period that begins at her arrival and ends at her departure, such as, e.g., the residual inter-arrival time at her departure. ■

In the remainder of this section we denote R_n and R_n^* , $n \geq 0$, by R and R^* , respectively. For its explicit formula we refer to Remark 4.1 at the end of this section. [8, 10] consider the similar residual, but then at arbitrary instances, and show that the independence in Theorem 3 extends only to where it is given that $n \geq 1$. Hence, two distributions are derived, one for the case where $n = 0$ and one for the case where $n \geq 1$. The following theorem presents a well known result for the steady state probabilities a_n and π_n , though the characterization of σ is new.

Theorem 4 *Let $\sigma = G^*(\mu)/(1 - R^*(\mu))$. Then,*

$$\frac{\pi_{n+1}}{\pi_n} = \frac{a_n}{a_{n-1}} = \sigma, \quad n \geq 1. \quad (9)$$

In particular, the above two ratios are not function of n as long as $n \geq 1$.

Proof: In order to initiate an n -two-step up transition, there must be a transition from state $n-1$ to state n , i.e., an arrival who finds $n-1$ customers in the system. Of course, such an arrival may be followed by a departure, and in that case an n -two-step up transition will not occur. Otherwise, if this arrival is followed by an additional arrival, this will indeed form an n -two-step up transition. Hence, the n -two-step up transition rate is equal to the rate of arrivals who find $n-1$ customers in the system, $a_{n-1}\lambda$, multiplied by the probability that such arrival will be followed by an additional arrival before an exponential service completion, which is $\int_0^\infty e^{-\mu x} dG(x) = G^*(\mu)$. In a similar way, the n -two-step down transition rate is equal to the rate of departures who leave behind n customers, that (by level crossing) equals to the rate of arrivals who find n customers in the system, $a_n\lambda$, multiplied by the probability that such departure will be followed by a consecutive departure before the next arrival. Theorem 3 implies that the distribution of the residual inter-arrival time at the moment of departure, is equal to R with LST R^* . Hence the probability that a consecutive exponential departure takes place before the next arrival equals $\int_0^\infty (1 - e^{-\mu x}) dR(x) = 1 - R^*(\mu)$. Summarizing, by RBP,

$$a_{n-1}\lambda G^*(\mu) = a_n\lambda(1 - R^*(\mu)), \quad n \geq 1, \quad (10)$$

which leads to

$$a_n/a_{n-1} = G^*(\mu)/(1 - R^*(\mu)), \quad n \geq 1. \quad (11)$$

This completes proof for the first part of theorem. The second part immediately follows by observing that the rate of arrivals who find $n - 1$ customers in the system, $a_{n-1}\lambda$, equals (by level crossing) to the rate of departures who leave behind $n - 1$ customers, $\pi_n\mu$. ■

The above theorem implies the following corollary (see, e.g., p. 100 in [7]).

Corollary 1 *The fact that $\pi_0 = 1 - \rho$ and (9) imply that*

$$a_n = (1 - \sigma)\sigma^n, \quad n \geq 0 \quad (12)$$

$$\pi_n = \rho(1 - \sigma)\sigma^{n-1}, \quad n \geq 1. \quad (13)$$

Moreover, the geometric distribution in (12) implies that under the FCFS service regime, the sojourn time is the sum of a geometric (random) number with parameter $1 - \sigma$ of i.i.d exponential variables with parameter μ , the distribution of which is exponential with rate $\mu(1 - \sigma)$.

The theorem below is the “usual” characterization of σ (see, e.g., p. 100 in [7]), for which we now suggest a new proof.

Theorem 5 *σ is the unique value obeying*

$$\sigma = G^*(\mu(1 - \sigma)) \quad (14)$$

and $0 < \sigma < 1$.

Proof: We show that the two sides of (14) are probabilities of the same event. The left hand side is the probability of the event that an arbitrary arrival finds a non-empty queue, i.e., $1 - a_0 = \sigma$. Now, assume w.l.g that the service regime is FCFS. In that case, the sojourn time is exponential with rate $\mu(1 - \sigma)$ (see Corollary 1). The event that an arbitrary arrival finds a non-empty queue is also equal to the event that the sojourn time of the previous arrival is greater than her inter-arrival time. The probability of this event equals $\int_0^\infty e^{-\mu(1-\sigma)x} dG(x) = G^*(\mu(1 - \sigma))$, which is the right hand side of (14). The uniqueness of the solution of (14) can be argued using standard convexity arguments (see e.g. p. 101 in [7]). ■

Theorem 6 *The steady-state CDF of the residual inter-arrival time at a departure instance equals $D_{\mu(1-\sigma),G}$.*

Proof: Assume w.l.g the FCFS service regime. Under this regime, the sojourn time is exponential with rate $\mu(1-\sigma)$ (see Corollary 1). Tag a customer and let S to be her sojourn time, A be the next inter-arrival time after her arrival, and R be the residual inter-arrival time at her departure. Of course, A and S are independent. If $S < A$, then $R = A - S$. Otherwise, if $S > A$, the memoryless property of S implies that the remaining sojourn time at the moment of the next arrival, $S - A$, is again distributed as S and that implies that the distribution of the residual at the moment of departure is the same as the distribution of R . The above and the law of total probability imply that,

$$P(R < x) = P(A - S < x | S < A)P(S < A) + P(R < x)P(S > A).$$

Solving for $P(R < x)$ gives

$$P(R < x) = P(A - S < x | S < A)$$

as required. ■

Remark 4.1 Theorem 6 implies that $R^*(s) = D_{\mu(1-\sigma),G}^*(s)$. Equations (2) and (14) implies that,

$$R^*(s) = \frac{\mu(G^*(s) - \sigma)}{\mu(1-\sigma) - s}. \quad (15)$$

In particular, $R^*(\mu) = 1 - \frac{G^*(\mu)}{\sigma}$, which coincides with the definition of σ in Theorem 4.

Remark 4.2 Observe that R_0 is the idle period in this model. Hence (15) gives the LTS of the idle period, which coincides with Theorem 1 in [2].

5 The G/Mn/1 queueing model

Here we deal with a variation of the G/M/1 model where the service rate is queue length dependent. More formally, given that there are n customers in the system at time t , the number of service completions within the time interval $[t, t + \Delta]$ is independent of the past. Moreover, the probability of a single service completion within this time interval equals $\mu_n \Delta + o(\Delta)$ and

the probability of two or more service completions equals $o(\Delta)$. The arrival process is as in the G/M/1 model. Except for the service rates, we use the same notations as done in the previous section.

In this section we use the RBP in order to show the following. First, we show that Theorem 4 can be generalized to the G/Mn/1 model, where μ_n replaces μ . Then, we show that the probability that a customer who leaves behind n customers upon departure, is the first to depart during the current inter-arrival period equals $1 - G^*(\mu_{n+1})$ (which is the probability that an inter-arrival period exceeds an exponential random variable with parameter μ_{n+1}). Finally, we derive an original recursion for the CDF of R_n , $n \geq 1$. In general, this recursion does not have easily computable initial conditions. Yet, we show that this can be overcome when μ_n becomes constant once n is large enough.

Theorem 7

$$\frac{\pi_{n+1}\mu_{n+1}}{\pi_n\mu_n} = \frac{a_n}{a_{n-1}} = \frac{G^*(\mu_n)}{1 - R_n^*(\mu_n)}, \quad n \geq 1. \quad (16)$$

Proof: The same arguments as in the proof of Theorem 4 yield that the rates of n -two-step up and down transitions, $n \geq 1$, equal $a_{n-1}\lambda G^*(\mu_n)$ and $a_n\lambda(1 - R^*(\mu_n))$, respectively. Equating these rates leads to the second equality in (16). Similarly, by level crossing we get that $a_{n-1}\lambda = \pi_n\mu_n$, yielding the first equality in (16). ■

Corollary 2 For $n \geq 1$,

$$a_n = a_0 \prod_{k=1}^n \frac{G^*(\mu_k)}{1 - R_k^*(\mu_k)}, \quad \pi_{n+1} = \pi_1 \frac{\mu_1}{\mu_{n+1}} \prod_{k=1}^n \frac{G^*(\mu_k)}{1 - R_k^*(\mu_k)}.$$

The next theorems give a recursion for R_n^* , $n \geq 0$, that is required in order to use the result of Corollary 2. The analysis here is a dual version of the analysis of the Mn/Gn/1 model in the next section.

Lemma 1 *The probability that a departure who leaves behind $n \geq 0$ customers in the system, is the first to depart during the current inter-arrival time equals $1 - G^*(\mu_{n+1})$. In particular, this probability equals the probability that an inter-arrival time who commences when $n + 1$ customers are in the system will end after completion of the customer currently in service.*

Proof: We use the same TST argument as in the proof of Theorem 7 (and Theorem 4). We already argued that the rate of $n+1$ -two-step up transitions

equals $a_n \lambda G^*(\mu_{n+1}) = \pi_{n+1} \mu_{n+1} G^*(\mu_{n+1})$, but the rate of $n+1$ -two-step down transitions can now be argued differently than before. An $n+1$ -two-step down transition occurs when and only when a departure leaves behind n customers in the system, and she is not the first to depart during the current inter-arrival time. Hence the rate of $n+1$ -two-step down transitions equals the rate of departures leaving behind n customers, $\pi_{n+1} \mu_{n+1}$, times the probability that such departure is not the first to depart during the current inter-arrival time, i.e., one minus the probability of being the first. Now, the result follows by equating the two-step up and down transition rates. ■

Remark 5.1 Alternative proof of Lemma 1

For each departure who leaves behind n customers in the system, there is an arrival who finds n customers in the system, or, in different words, an inter-arrival time that was initiated with $n+1$ customers. Hence, the rate of such departures, denoted by δ_n , and the rate of such arrivals, denoted by α_n , are equal. Furthermore, for each departure who leaves behind n customers in the system and is the first to depart during the current inter-arrival time, there is an inter-arrival time that was initiated with $n+1$ customers, and had at least one departure during it. Hence, the rate of such first departures, denoted by δ_n^f , and the rate of such inter-arrival times, denoted by α_n^+ , are equal. The probability we are after equals the proportion of the rate of departures who leave behind n and are first to depart out of the total rate of departures who leave behind n customers. In the above notation it equals to δ_n^f / δ_n , and the explanation above implies that it equals to α_n^+ / α_n . The last term equals the probability that an inter-arrival time that was initiated with $n+1$ customers in the system, will have at least one departure during it. This probability equals the probability that a generic random variable with CDF G is greater than an exponential random variable with rate μ_{n+1} , which is equal to $\int_0^\infty (1 - e^{-\mu_{n+1}x}) dG(x) = 1 - G^*(\mu_{n+1})$.

Theorem 8

$$R_n^*(s) = (1 - G^*(\mu_{n+1})) D_{\mu_{n+1}, G}^*(s) + G^*(\mu_{n+1}) D_{\mu_{n+1}, R_{n+1}}^*(s), \quad n \geq 0 \quad (17)$$

Proof: Consider a departure who leaves behind n customers in the system and is the first to depart during the current inter-arrival time. The fact that this customer is the first to depart implies that the distribution of the residual inter-arrival is $D_{\mu_{n+1}, G}$. Now, consider an arrival who leaves behind n customers in the system and is not the first to depart during the current

inter-arrival time. That means that there is a customer that departed before her, and left behind $n + 1$ customers. From the moment of the previous departure, a residual inter-arrival with distribution of R_{n+1} was initiated, and the current departure is the first to depart during it. Hence, in this case, the distribution of the residual inter-arrival time is $D_{\mu_{n+1}, R_{n+1}}$. Finally, using the law of total probability along with the probability in Lemma 1 completes the proof. \blacksquare

Remark 5.2 The recursion in (17) can be used in both directions. Of course, if R_{n+1}^* is in hand, one can apply the recursion to get R_n^* . Likewise, if R_n^* is in hand, one can solve (17) for $D_{\mu_n, R_{n+1}}^*$ and use (8) to get R_{n+1}^* .

The above theorem gives a recursion, but does not specify a starting point, i.e., R_k for some $k \geq 0$ that can be used in order to apply (17) recursively. Unfortunately, for the general case, such starting point is not available, but the following theorem implies that the complexity of the computation of each R_n , $n \geq 0$, does not depend on n .

Theorem 9 *Let R_n , $n \geq 0$ be the CDF of the conditional residual inter-arrival time at departure instance in a $G/Mn/1$ queueing model with service rates μ_n . For $k \geq 0$, let $R_n^{(k)}$, $n \geq 0$, be the CDF of the residual inter-arrival time at departure instance, conditioned on queue length n , associated with a $G/Mn/1$ model with service rates $\mu_n^{(k)}$, such that $\mu_n^{(k)} = \mu_{n+k}$, $n \geq 1$. Then,*

$$R_n^{(k-m)} = R_0^{(k)}, \quad 0 \leq m \leq k. \quad (18)$$

In particular,

$$R_k = R_0^{(k)}, \quad k \geq 0.$$

Proof: We use a similar argument as in the proof of Theorem 3. Consider the system associated with service rates $\mu_n^{(k-m)}$, for some $0 \leq m \leq k$. Assume w.l.g that the LCFS-PR service regime is used. The residual inter-arrival time at the departure of a customer who leaves behind m customers is a function of the service process only from her arrival to her departure. This process is stochastically equivalent to the process of a customer who finds 0 customers in a system with service rates $\mu_n^{(k)}$. \blacksquare

Corollary 3 *For any fixed $k \geq 0$,*

$$R_{k+m} = R_m^{(k)}, \quad m \geq 0.$$

Moreover, (16) implies that the steady state distribution of the difference between the number of customers in the system and k , conditioned on the

queue length being greater than or equal to k , equals the steady state queue length distribution in the $G/Mn/1$ model with service rates $\mu_n^{(k)}$.

Remark 5.3 Consider the special case, where from some (arbitrary large) queue length, service rates are equal. Specifically, assume that there exists an $N \geq 1$ and $\bar{\mu}$ such that $\mu_n = \bar{\mu}$, for all $n \geq N$. An immediate consequence of Corollary 3 is that for $n \geq N$, $R_n = \bar{R}$, where \bar{R} is the steady state residual inter-arrival time distribution in a $G/M/1$ model with constant service rate $\bar{\mu}$. This means that, starting with R_N^* from (15), the transforms $R_{N-1}^*, R_{N-2}^*, \dots, R_0^*$ can be computed recursively using (17).

Example 5.1 $G/M/c$ model

Consider the $G/M/c$ model. This model is probabilistically equivalent to the $G/Mn/1$ model with $\mu_n = n\mu$, $1 \leq n \leq c$, and $\mu_n = c\mu$, $n > c$. Using the result in Theorem 9 and (15) we get that

$$R_c^*(s) = \frac{c\mu(G^*(s) - \sigma)}{c\mu(1 - \sigma) - s},$$

where σ is the unique solution of $\sigma = G^*(c\mu(1 - \sigma))$ and $0 < \sigma < 1$. Using the results from Corollaries 3 and 1 we get that

$$\pi_n = C\sigma^n, \quad n \geq c$$

for some $C > 0$. The remaining probabilities, π_n , $0 \leq n < c$, can be computed using (16) and (17). Observe that this method does not require to compute infinite sums, as in the embedded Markov chain analysis of this model, see e.g. p. 348 in [3].

6 The $Mn/Gn/1$ queueing model

The model dealt with in this section is similar to the model analyzed in [9], with the addition that service time distribution is state dependent. Consider a single server queueing system where service times are independent but not necessarily identically distributed. The distribution of a service time depends on the state of the system upon service commencement. Specifically, the distribution of a service time that commence with $n \geq 1$ customers in the system is with CDF G_n and LST G_n^* . The arrival process is as follows. Given that there are n customers in the system at time t , the number of arrivals within the time interval $[t, t + \Delta]$ is independent of the past. The

probability of a single arrival within this time interval equals $\lambda_n \Delta + o(\Delta)$ and the probability of two or more arrivals equals $o(\Delta)$.

Assume that the system is stable. Let π_n , $n \geq 0$, be the steady state probabilities of having n customers in the system at arbitrary instance, and let R_n and R_n^* , $n \geq 1$ be the CDF and LST, respectively, of the residual service time at an arrival instance, given that there are n customers in the system.

We start this section by a recursion for π_n , which does not have easily computable initial conditions. We then show that for $n \geq 2$, an arrival, who sees n customers upon his arrival, is with probability $1 - G^*(\mu_n)$ the first to appear during the current service period. We end with a recursion on R_n , $n \geq 1$, and provide simple initial conditions.

The above model is similar to the Mn/Gn/1 model in [1], except that service rates in [1] are not constant, but can change at arrival instances. In [1] the steady state probabilities and a recursion for the residual conditional service time are derived using the method of supplementary variables [4]. We derive these results for our related model (though the extension to the model in [1] is straightforward) by using probabilistic and RBP arguments, similar to the ones in the previous sections. We further like to point out that [5] also uses probabilistic arguments (but different from the ones used below) to derive the results in [9] for the Mn/G/1 model.

Theorem 10 *The steady state probabilities for the Mn/Gn/1 model obey the following birth and death like equations:*

$$\pi_{n-1} \lambda_{n-1} (1 - R_{n-1}^*(\lambda_n)) = \pi_n \lambda_n G_n^*(\lambda_n), \quad n \geq 1 \quad (19)$$

where $R_0^* = G_1^*$.

Proof: We use TST as in the analysis of the G/M/1 model. In order to initiate an n -two-step up transition, there must be a transition from state $n-1$ to state n , i.e., an arrival who finds $n-1$ customers in the system. This arrival should be followed by an additional arrival in order for an n -two-step arrival to occur. Hence, the n -two-step up transition rate is equal to the rate of arrivals who find $n-1$ customers in the system, $\pi_{n-1} \lambda_{n-1}$, multiplied by the probability that such arrival will be followed by an additional arrival. At the moment of the first arrival, for $n > 1$, a residual service time is initiated, and for $n = 1$ a fresh service is initiated. The probability of the event that an additional arrival occurs before this residual (or fresh) service time equals the probability that a generic random variable with distribution R_{n-1} (or $R_0 = G_1$, respectively) is greater than an exponential random variable with

rate λ_n , given by $1 - R_{n-1}^*(\lambda_n)$. Similarly, in order to initiate an n -two-step down transition, there must be a transition from state $n + 1$ to state n , i.e., a departure who left behind n customers in the system. Of course, such a departure must be followed by a consecutive departure in order to form an n -two-step down transition. Hence, the n -two-step down transition rate is equal to the rate of departures who leave behind n customers, that equals to the rate of arrivals who find n customers in the system, $\pi_n \lambda_n$, multiplied by the probability that such departure will be followed by a consecutive departure before the next arrival. At the moment of the first departure, a fresh service time is initiated. Hence, the probability of a second departure equals the probability that a generic random variable with distribution G_n is less than an exponential random variable with rate λ_n . This probability equals $G_n^*(\lambda_n)$. ■

Corollary 4 For $n \geq 1$,

$$\pi_n = \pi_0 \frac{\lambda_0}{\lambda_n} \prod_{k=1}^n \frac{1 - R_{k-1}^*(\lambda_k)}{G_k^*(\lambda_k)}.$$

In order to use Corollary 4 for calculation of the steady-state probabilities, the LSTs of the conditional residual service times are required. These LSTs can be derived using the method in [9] for the Mn/G/1 model, but we propose an alternative probabilistic approach.

Lemma 2 *The probability that an arrival who finds $n \geq 2$ customers in the system, is the first to arrive during the current service equals $1 - G_n^*(\lambda_n)$. In particular, this probability equals the probability that a service who commences when n customers are in the system will be completed after the next arrival.*

Proof: We use the same TST argument as in the proof of Theorem 10. We already argued that the rate of n -two-step down transitions equals $\pi_n \lambda_n G_n^*(\lambda_n)$, but the rate of n -two-step up transitions can now be argued differently than before. An n -two-step up transition occurs when and only when an arrival finds n customers in the system, and she is not the first to arrive during the current service. Hence, the rate of n -two-step up transitions equals the rate of arrivals who find n customers, $\pi_n \lambda_n$, times the probability that such arrival is not the first to arrive during the current service time, i.e., one minus the probability of being the first. Now, the result follows by equating the up and down transition rates. ■

Remark 6.1 Alternative proof of Lemma 2

Just as in Remark 5.1, an alternative proof can be given by calculating the proportion of the rate of arrivals who find n and are first to arrive out of the total rate of arrivals who find n customers upon arrival.

Theorem 11 *The conditional residual service time distributions R_n , $n \geq 1$, follow the recursion*

$$R_1 = D_{\lambda_1, G_1} \quad (20)$$

and

$$R_n = (1 - G_n^*(\lambda_n))D_{\lambda_n, G_n} + G_n^*(\lambda_n)D_{\lambda_n, R_{n-1}}, \quad n \geq 2. \quad (21)$$

Proof: Consider an arrival who finds n customers in the system and is the first to arrive during the current service. The fact that this customer is the first to arrive implies that the distribution of the residual service time is equal to D_{λ_n, G_n} . Now, Consider an arrival who sees n customers in the system and is not the first to arrive during the current service. That means that there is a customer that arrived before her, and saw $n - 1$ customers. From the moment of the previous arrival, the residual service time is distributed as R_{n-1} , and the current arrival is the first to arrive during this residual service time. Hence, the distribution of the residual service time at the current arrival is $D_{\lambda_n, R_{n-1}}$. Finally, use of the law of total probability along with the probability in Lemma 2 completes the proof. ■

7 Summary

In this paper we introduced a rate balance principle for general stochastic processes. For the special case of birth and death like processes, we showed that it leads to a new balance principle between the rates of two consecutive up transitions and of two consecutive down transitions, and demonstrated the potential of this “two step transition” rate principle to probabilistically derive some well known, but also new results for the G/Mn/1 and Mn/Gn/1 queues, such as recursions on the conditional distributions of the residual service and inter-arrival times, given the queue length. Hence, we believe that this principle is a promising tool to also explore other stochastic processes.

Acknowledgement

This research was partly supported by Israel Science Foundation grant no. 1319/11.

References

- [1] Abouee-Mehrizi, H. and O. Baron (2015), “State-Dependent M/G/1 Queueing Systems,” *Queueing Systems*, (to appear).
- [2] Adan, I., Boxma O., and D. Perry (2005), “The G/M/1 queue revisited,” *Mathematical Methods of Operations Research*, **62**, 437–452.
- [3] Asmussen, S. (2003) *Applied Probability and Queues*, Springer.
- [4] Cox, D. R. (1955), “Use of Complex Probabilities in the Theory of Stochastic Processes,” *Proceedings of the Cambridge Philosophical Society*, **51**, 313–319.
- [5] Economou, A. and A. Manou (2015), “A probabilistic approach for the analysis of the Mn/G/1 queue,” *Annals of Operations Research*, Advance online publication, doi:10.1007/s10479-015-1943-0.
- [6] El-Taha, M. and S. Stidham Jr. (1999) *Sample-Path Analysis of Queueing Systems*, Springer.
- [7] Haviv, M. (2013) *Queues - A Course in Queueing Theory*, Springer.
- [8] Haviv, M. and Y. Kerner (2011), “The age of the arrival process in the G/M/1 and M/G/1 queues,” *Mathematical Methods in Operations Research*, **73**, 139–152.
- [9] Kerner, Y. (2008), “The conditional distribution of the residual service time in the Mn/G/1 queue,” *Stochastic Models*, **24**, 364–375.
- [10] Nunez-Queija, R. (2001), “Note on the GI/GI/1 queue with LCFS-PR observed at arbitrary times,” *Probability in the Engineering and Informational Sciences*, **15**, 179–187.